Defining operational taxonomic units using DNA barcode data

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2005) **360**, 1935–1943
doi:10.1098/rstb.2005.1725
Published online 12 September 2005

Defining operational taxonomic units using DNA barcode data

Mark Blaxter*, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd† and Eyualem Abebe‡

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, King's Buildings, Edinburgh EH9 3JT, UK

The scale of diversity of life on this planet is a significant challenge for any scientific programme hoping to produce a complete catalogue, whatever means is used. For DNA barcoding studies, this difficulty is compounded by the realization that any chosen barcode sequence is not the gene 'for' speciation and that taxa have evolutionary histories. How are we to disentangle the confounding effects of reticulate population genetic processes? Using the DNA barcode data from meiofaunal surveys, here we discuss the benefits of treating the taxa defined by barcodes without reference to their correspondence to 'species', and suggest that using this non-idealist approach facilitates access to taxon groups that are not accessible to other methods of enumeration and classification. Major issues remain, in particular the methodologies for taxon discrimination in DNA barcode data.

Keywords: DNA barcodes; molecular operational taxonomic units; tardigrades; nematodes; meiofauna; small subunit ribosomal RNA

1. INTRODUCTION: THE UNSEEABLE ANIMAL

The total number of unique taxa described to the species level is circa 1.5 million, but the total number of 'species' is likely to be in the region of 10 million (May 1988). The overall 'taxonomic deficit' (the ratio of expected taxa to named taxa) is thus approximately sixfold. However this deficit, like all phylogenetic things, is not immune to systematic bias (Blaxter 2003). For vertebrates, the current described species total is likely to be relatively close to the 'true' total: we have described most of these relatively large organisms. The same is true of most groups whose members have body sizes greater than 10 mm. However, the vast majority of organisms on the Earth have body sizes less than 1 mm, and for these groups the taxonomic deficit is likely to be several fold worse than for land plants and vertebrates (Lambshead 1993; Platt 1994; Lambshead & Boucher 2003). These meio- and micro-fauna and flora are, however, key to the functioning of ecosystems and are the productive and saprophytic base upon which the macro-organisms rely. Their size precludes facile visual identification, and indeed much of their important morphology may be at scales that are beyond the resolution of light microscopy (De Ley & Bert 2001; De Ley *et al.* 2005). Wendell Berry quotes from his daughter in his poem 'To the unseeable animal': *'I hope there's an animal somewhere that nobody has ever seen./ And I hope nobody ever sees it.'* (Berry 1970). We suggest that DNA barcoding may permit rational access to these animals.

DNA barcoding, the use of a specified DNA sequence to provide taxonomic identification for a specimen, is a technique that should be applicable to all cellular (and much viral) life (Floyd *et al.* 2002; Hebert *et al.* 2003; Tautz *et al.* 2003; Blaxter *et al.* 2004). Theoretically, this should allow rapid and high-throughput identification, either of individual organisms or of sequences isolated from an environmental DNA sample. Specimen-independent DNA surveys are already used for microbial (Giovannoni *et al.* 1990) and protozoal communities (Diez *et al.* 2001; Lopez-Garcia *et al.* 2001; Moreira & Lopez-Garcia 2002; Amaral Zettler *et al.* 2002), and have revealed a wealth of hidden diversity. Meiofauna would appear to be an ideal group in which a molecular identification system could be used (Lambshead 1993; Lawton *et al.* 1998; Blaxter 2004).

2. BARCODING MEIOFAUNA: CHALLENGES

The number of meiofaunal taxa, animals with a body size ~1 mm (or less), can only be guessed at. Thus, the number of described species of nematodes is quoted as between 26 000 and 40 000, but the real total estimated to be above one million (Lambshead 1993; Platt 1994; Lambshead & Boucher 2003). The deficit may be put into perspective by considering that the number of described species of soil dwelling nematodes for the UK is approximately 400, a figure surprisingly close to the inventory of UK breeding birds. Is the UK nematode fauna really that depauperate? Our surveys of nematodes in soils in relatively degraded habitats (upland farm grassland) suggest that taxon numbers identifiable from even a small area may be remarkably high (R. Floyd, A. Eyualem and M. Blaxter,

* Author for correspondence (mark.blaxter@ed.ac.uk).
† Present address: British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK.
‡ Present address: Department of Zoology, Hubbard Center for Genome Studies, University of New Hampshire, Durham NH 03824, USA.

One contribution of 18 to a Theme Issue 'DNA barcoding of life'.

unpublished). Similarly, for tardigrades, the described UK fauna is ~100 species (Maucci 1986; Kinchin 1994), but we have identified over 50 taxa from one restricted set of sample sites (Blaxter *et al.* 2003). While some authors have argued for a relatively low number of meio-taxa matched by a near-ubiquitous distribution (Finlay 2002), we have found that different sites, though close geographically, can have very different taxon assemblages (Blaxter *et al.* 2003). If organisms with a body size <1 mm really do have no biogeographical structure, and are all essentially ubiquitous, the sampling we have carried out suggests at least that relative abundances must vary greatly between sites. Meiofaunal barcoding must fall into the purview of the third community identified above: experimental investigation of biodiversity.

We have been generating DNA barcode datasets for meiofaunal specimens (mostly nematodes and tardigrades) for several years (Floyd *et al.* 2002; Blaxter & Floyd 2003; Blaxter *et al.* 2003; Eyualem & Blaxter 2003; Blaxter 2004; Blaxter *et al.* 2004). We are agnostic as to whether the taxa we can define using these barcode sequences (which we call 'molecular operational taxonomic units' or MOTU) are 'species' or not, though in the case where we have compared and contrasted MOTU, morphological species hypotheses and breeding-based biological species, MOTU and biological species hypotheses were congruent while morphological analyses disagreed internally, and with the other modes of taxon definition (Eyualem & Blaxter 2003). We have traditionally used the nuclear small subunit (nSSU) as a marker, but have also tested nSSU alongside cytochrome oxidase subunit I (*cox1*), with equivalent resolution. Here, we use a new dataset of meiofaunal barcodes to discuss what we feel are very interesting and important features of DNA barcode data: they can be used not only to define taxa, but also to identify sets of specimens for which robust taxonomic hypotheses are difficult to construct. These clouds of related specimens are immediately of interest for further study: is this evidence for recent, rapid radiation of distinct taxa or is it evidence for a highly variable single taxon?

3. METHODS: OBTAINING MEIOFAUNAL BARCODE SEQUENCES

(a) *Sampling of moss ecosystems*

Moss samples for this study were collected from dry stone walls surrounding Ettrick Old Church, in Glen Ettrick in Southern Scotland (Blaxter *et al.* 2003). Meiofauna were isolated by modified Baermann funnel separation through milk filters into sterile tap water. Larger fauna (such as collembolans and mites; body sizes >2 mm) were excluded from the separation by the pore size of the filter: some of these arthropods were picked from moss individually. Relative numbers of animals from each phylum were counted from a subsample of the filtrate, and a few of each phylum picked individually: the remainder was processed for DNA extraction.

(b) *Individual specimen barcoding*

Individual animals were extracted using the NaOH direct lysis procedure: this yields ~40 µl of stable extract per specimen from which over a dozen PCRs can be performed (Floyd *et al.* 2002). Bulk filtrate animals were concentrated by centrifugation and extracted using a snap-freezing/proteinase K/phenol/chloroform protocol. The nSSU marker was amplified from individual extracts using the primers SSU_F04 (GCTTGTCTCAAAGATTAAGCC) and SSU_R26 (CATTCTTGGCAAATGCTTTCG) (Blaxter *et al.* 1998), yielding a ~900 base pair (bp) product. These primers were designed to be metazoan-specific (Blaxter *et al.* 1998). The *cox1* amplicons were amplified from a subset of tardigrade individuals (also amplified for nSSU) using the 'universal' primers *cox1* (HC02198; TAAACTTCAGGGTGACCAAAAAATCA) and *cox1* (LC01490; GGTCAACAAATCATAAAGATATTGG) (Hebert *et al.* 2003), yielding a ~650 bp product. Shrimp alkaline phosphatase/exonuclease I-cleaned PCR products from single specimens were sequenced directly using SSU_R09 or *cox1*.

(c) *Barcodes from bulk ecosystem DNA*

Bulk filtrate animals were concentrated by centrifugation and extracted using a snap-freezing/proteinase K/phenol/chloroform protocol. The nSSU marker was amplified as described above. Amplicons of nSSU generated from the bulk extract target were cleaned using a Montage gel extraction kit and cloned into pTOPO2.1 (Invitrogen). After growth on LB/kanamycin/IPTG/Xgal, recombinant colonies were picked to 200 µl of LB broth with kanamycin in microtitre plates and grown overnight. Inserts in the recombinant plasmids were amplified from ~1 µl of overnight liquid culture using the primers M13_F (CTGGCCGTCGTTTAC) and M13_R (CAGGAAACAGCTATA), cleaned using shrimp alkaline phosphatase and exonuclease I, and sequenced using SSU_R09 (AGCTGGAATTACCGCGGCTG) and ABI BigDye3.0 reagents to produce ~500 bp of sequence.

(d) *Molecular operational taxonomic unit definition*

The sequencing was carried out on an ABI3730 capillary sequencer, and sequencing chromatograms were post processed with trace2seq (a perl program that uses phred to identify high-quality base calls and crossmatch to identify vector sequence; A. Anthony and M. Blaxter, unpublished). All sequences have been deposited in EMBL/GenBank/DDBJ. The perl program 'MOTU_define.pl' (R. Floyd and M. Blaxter, unpublished; based on CLOBB (Parkinson *et al.* 2002)) was used to allocate the resulting high-quality sequences to MOTU, based on pairwise identity scores and a user-defined cutoff.

The MOTU_define.pl program adds sequences one at a time to a growing database of barcode sequences (figure 1). It is a very simple procedure, internally consistent, and has the benefit of allocating stable MOTU identifiers to the dataset. As more sequences are generated, they can be added incrementally to the existing MOTU sets and thus continuity between experiments is attained. Indeed, sequence data can be acquired from other sources (such as GenBank/EMBL) and added to the dataset without compromising or changing the MOTU

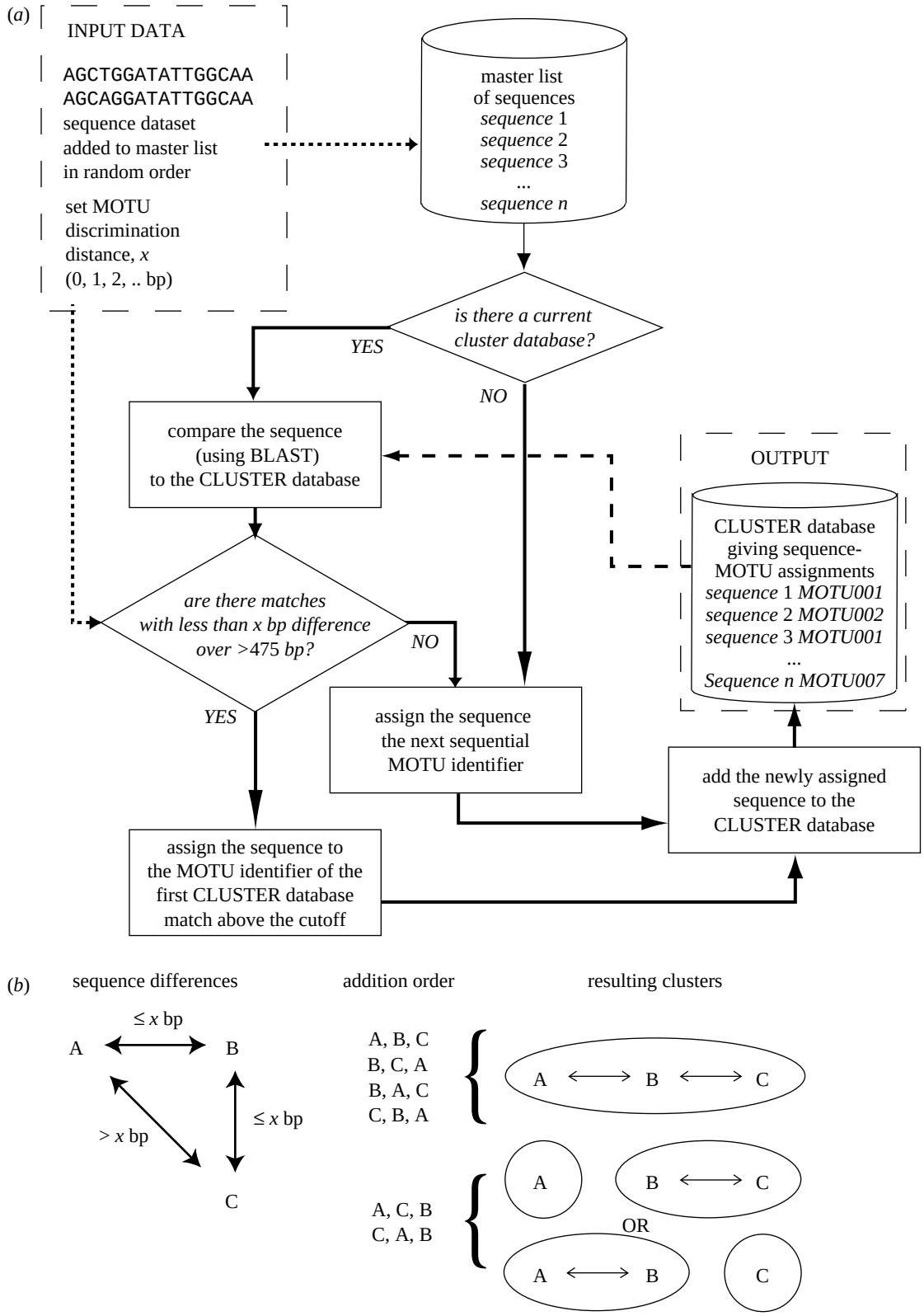

Figure 1. The MOTU_define.pl system. (*a*) A schematic of the process by which MOTU_define.pl allocates sequences to MOTU. The process can be run any number of times with different sequence addition order to assess MOTU stability. (*b*) The effect of addition order on MOTU definition. Three sequences, A, B and C, are clustered into MOTU. A differs from B, and B from C by less than the MOTU discriminant cutoff, but C differs from A by more than the cutoff. Depending on the order of analysis of the sequences, either one or two MOTU will be defined.

assignment of local data. This sort of process is ideal for building up a shared database of MOTU assignments and sequences. It is relatively rapid and reasonably scaleable (a variant of the program using the megaBLAST algorithm can cluster 100 000 expressed sequence tag sequences in ~20 h on a desktop computer (Parkinson *et al.* 2002); we do not yet have barcode datasets of this magnitude to test). The MOTU_define.pl is freely available from M. Blaxter, and requires only perl and a local copy of the NCBI BLAST suite (it is thus installable on UNIX, MacOSX and Windows systems).

Following our previous analyses of similar data, and our measured error rate in sequencing (~1 base in 3500) (Floyd *et al*. 2002), we standardly use a cutoff of 2 base differences in ~500 bp of sequence to discriminate MOTU: this can be varied. The program can also be rerun multiple times over the same set of sequences, randomizing the input order each time, and thus can be used to identify sequences and MOTU that do not behave simply under the cutoff statistic used. The use of single linkage clustering in MOTU_define.pl (where each sequence is clustered based on its identity to a single comparator) avoids issues of ambiguous alignment across a wide range of distantly related sequences. The high-quality sequences were aligned to each other and to a set of relevant control sequences from named taxa derived from GenBank or our previous studies and the alignment analysed using Maximum Parsimony in PAUP* 4.0b10 (Swofford 1999).

4. RESULTS: ETTRICK MOSS MEIOFAUNA

The moss fauna included animals from five animal phyla: Arthropoda (mites and collembolans), Tardigrada, Annelida (enchytraeids), Nematoda and Rotifera. The filtrates also included many protozoa (ciliates and amoebas) and some plant material. There was doubtless also a thriving unicellular fungal and algal, and prokaryote presence. Nematodes were most abundant, followed by rotifers and tardigrades (a ratio of 132 ± 20.8 nematodes to 6 ± 0.6 rotifers to 3 ± 0.9 tardigrades; mean and standard error of four samples corresponding to 0.5% of the extract from ~1 g dry weight of moss ecosystem); collembolans, mites and enchytraeids were rare in the moss, and excluded by their size from the filtrate.

(a) *Barcode sequence generation from single specimens*

Barcode sequences were derived from single specimens of nematodes, mites, collembolans, and enchytraeids. A total of 121 *cox1* sequences were generated from over 270 tardigrade specimens. For all taxa except rotifers, nSSU PCR and sequencing was successful ~85% of the time. In contrast, the *cox1* success rate was less than 40%. Indeed examination of available *cox1* sequences from animals related to those expected to be found in the moss ecosystem revealed that the 'universal' primers employed were unlikely to be able to amplify from some phyla. We conclude that use of the *cox1* target for the full diversity of animals will require additional rounds of primer pair optimization. No PCRs were successful from individually extracted rotifers, despite the nSSU primers sites being present in the available rotifer nSSU sequences. While this result could be due to the low number of cells (and thus genomes) in an individual rotifer, sequences from the bulk DNA sample were also rotifer-free (see below). We conclude that we will have to improve our extractions specifically to enhance rotifer DNA recovery.

(b) *Barcode sequences from nSSU libraries from bulk DNA*

A total of 145 sequences were generated from the bulk nSSU PCR library. Comparison to database sequences and single-specimen sequences from the same collection site (Blaxter *et al*. 2003) indicated that most derived from nematodes (123 or 85%) and four from tardigrades (3%). This ratio corresponds to that derived from the visual survey, excepting that no rotifer nSSU was recovered. In addition to these animal sequences, we isolated 18 nSSU sequences that clearly derived from ciliate protozoa, though none had an exact match in the public databases. We presume that these DNA segments were amplified because our primer set is not strictly metazoan-specific (we know that we can amplify environmental fungi, data not shown) and because, despite their being unicellular protozoa, ciliate macronuclei contain a many thousand fold amplification of the genes archived in the micronucleus, including the ribosomal RNA operons. No enchytraeid or arthropod sequences were recovered because the filtration excludes these larger meiofauna. Chimaeric amplicons are the bane of environmental sampling PCR. They arise from mispriming by amplification products during PCR, and result in DNA sequences that match one taxon at the 5′ end and another, unrelated one at the 3′ end. No chimaeric amplicons were identified, based on finding no discrepant BLAST matches for the first 250 compared to the last 250 bases of each.

(c) *Comparing single specimen and bulk nSSU MOTU*

MOTU_define.pl was used to infer MOTU from the nSSU datasets using a 2 bp difference cutoff. Data from the bulk sample and the single specimen sequences were clustered independently. For each nSSU MOTU, we derived a consensus sequence to represent that cluster for subsequent phylogenetic analysis (figure 2; but note that the definition of membership of a MOTU is not based on phylogenetic analysis). The use of a consensus sequence does not imply that this sequence correctly represents some ideal version of the true sequence, but rather is used to represent the diversity of the constituent sequences. The most abundant nSSU MOTU, derived from the bulk dataset, has 106 representatives, and is most similar to the chromadorid nematode *Plectus aquitilis*. Two of 16 single-specimen nSSU MOTU were also found in the bulk sample data (the *P. aquitilis*-like MOTU and a *Clarkus* (nematode)-like MOTU; figure 2). The bulk sequence dataset reflects the expected distribution of animals observed, excepting the Rotifera, and comparison with other more extensive datasets from soils and moss environments affirms that within the phyla that were amplified there is no apparent phylogenetic bias. Examination of this dataset suggests that the rate of identification of novel taxa using the barcode is not yet at saturation, despite the presence of the hyperabundant *P. aquitilis*-like Bulk_2bp_MOTU0001/Sin_2bp_ MOTU0005 (58% of all sequences, and 73% of the bulk sample sequences). Presumably, the rate of new MOTU identification could now be enhanced by prescreening for *P. aquitilis*-like sequences.

(d) *Comparison of cox1* and nSSU barcode analyses

A representative MOTU definition set for the *cox1* sequences is shown in figure 3. Twenty-two MOTU were defined, containing from 1 to 65 sequences

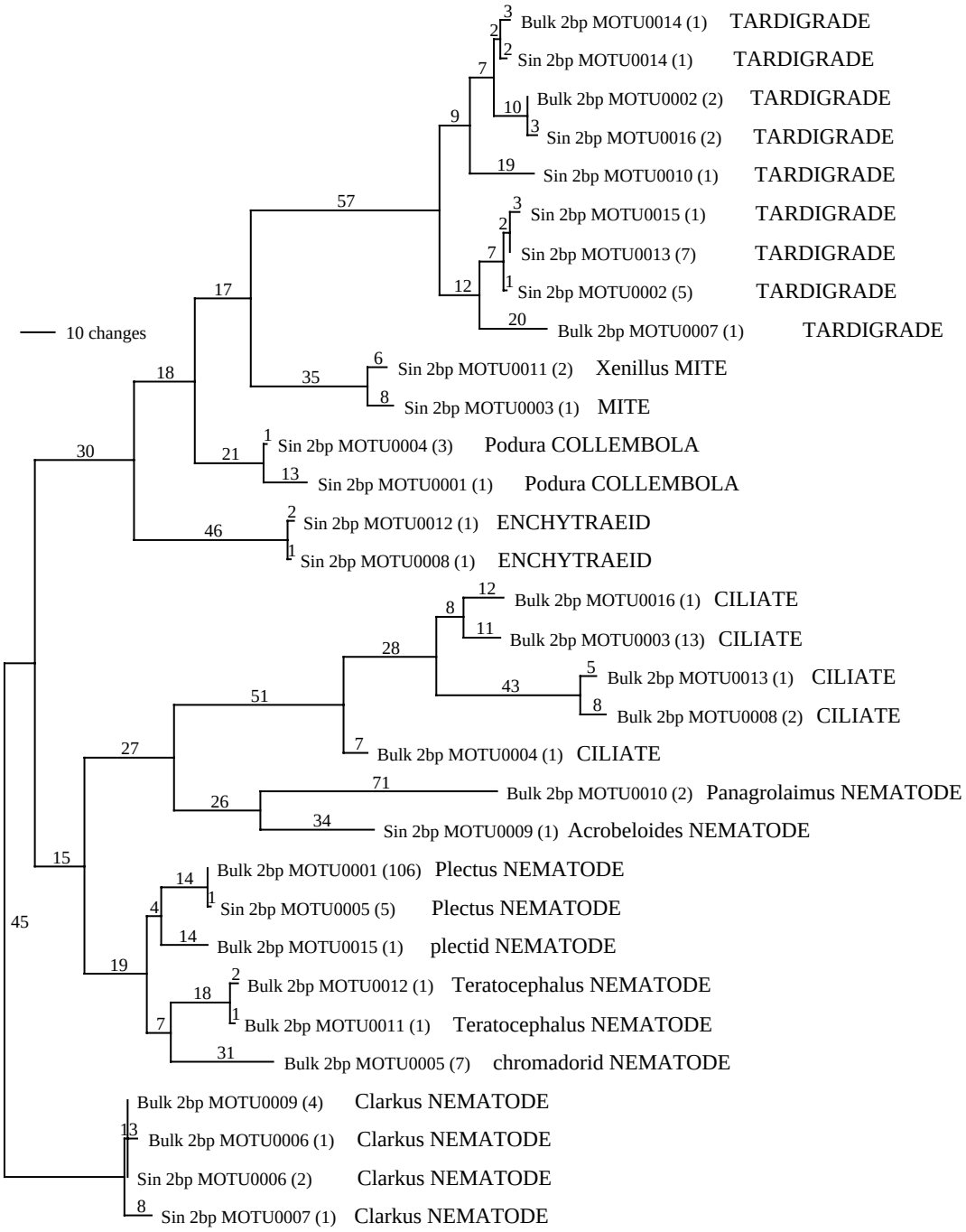

Figure 2. Meiofaunal MOTU defined using nuclear SSU sequences. The bulk nSSU dataset (Bulk) and a corresponding single specimen dataset (Sin) from the same moss sample were clustered into MOTU separately using a 2 bp cutoff, and consensus sequences predicted for those MOTU with more than one member. The consensus sequences and the singleton MOTU sequences were aligned and analysed using parsimony. For each MOTU represented the number of constituent sequences is given in brackets, and the taxonomic assignment based on BLAST search similarity to database sequences is given in bold. Where a taxon is identified below the major group, the MOTU sequence nested within a clade of sequences with the more specific designation (data not shown). Inferred numbers of changes are shown above each branch. Note that the tree is unrooted.

(figure 3*b*). The distribution of abundances of taxa implies one abundant taxon (~50% of the sample) and a larger number of taxa with low abundance.

For 82 tardigrade specimens, we obtained sequences of both *cox1* and nSSU with >490 bp of high-quality data. The two markers were used to infer independent clusterings, using a 2 bp cutoff, and the resultant clusters compared (figure 4). Seventeen *cox1* MOTU were defined from this subset. Surprisingly, 23 nSSU MOTU were defined, despite the overall lower level of sequence divergence, though the distance between distinct clusters was greater in the *cox1* dataset (as would be expected from the known higher substitution rate in animal mitochondrial genes). Seven MOTU with single members were found in both datasets, and two *cox1* MOTU (with two and five members) corresponded to two nSSU MOTU each (figure 4). The remaining 68 specimens formed two groups with complex patterns of overlap between nSSU and *cox1* MOTU (figure 4). Thus, while *cox1*

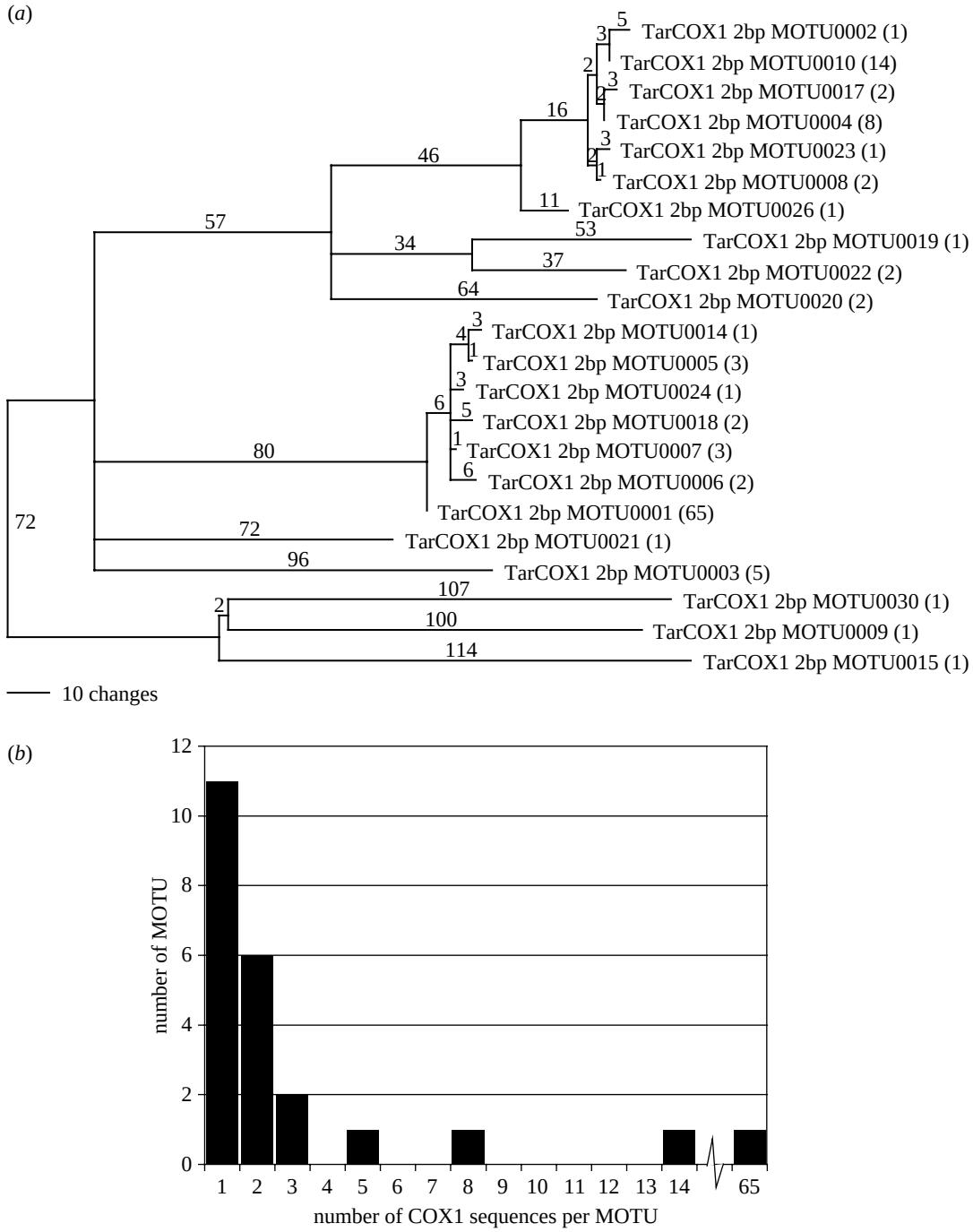

Figure 3. Tardigrade MOTU defined using *cox1* sequences. (*a*) A consensus sequence was derived for each MOTU, and these were aligned. The branch lengths are proportional to the number of discrete changes mapped to each. The number of sequences assigned to each MOTU is given in brackets after the MOTU name. (*b*) Histogram of MOTU abundance in the 121-sequence *cox1* dataset.

and nSSU are both effective at defining MOTU, and there was a general agreement between the two cluster sets, there were also significant disagreements. Whether these disagreements are due to the population history and hybridization patterns of the specimens sampled or are indicators of real incongruence between the markers is not clear. The two clouds of taxa (marked in figure 4) may correspond to particularly variable single taxa, or perhaps diverging radiations of taxa. Many tardigrades can reproduce asexually, or have sex only very rarely (Kinchin 1994), and thus this pattern may reflect divergence of clonal or matrilineal lines.

5. RESULTS: PROPERTIES OF EXACT SCORE MOTU DEFINITION

(a) *Variability due to single linkage clustering*

Assignment of any single sequence to a MOTU depends critically on what sequences have been added previously (figure 1*b*). If one takes three sequences, where only two differ by more than the chosen cutoff, the order of addition changes the number and membership of MOTU inferred. Rather than being a failing of this procedure, we regard this as being a feature: it permits exploration of the 'clouds' of taxa that are closely related. If a set of specimens robustly clusters into a particular set of MOTU, no

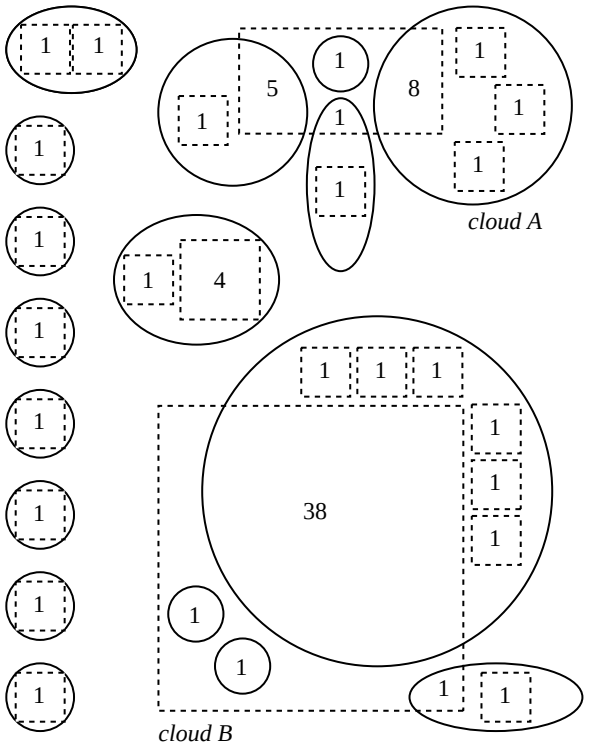

Figure 4. Comparison of MOTU definition using nSSU and *cox1* markers. This Venn diagram shows *cox1* MOTU sets (solid circles) and nSSU MOTU sets (dotted squares). The numbers within each partition indicate the number of individual specimens (out of 82) placed there.

matter what the addition order, this suggests that these MOTU have some congruence with biological taxa, and a distinctness from other related OTU. But if repeated clustering of a group of sequences yields discordant MOTU, this identifies a biologically interesting phenomenon, somewhere along the spectrum from population genetic processes to recently separated taxa still sharing ancestral polymorphisms. Such variability can thus alert researchers to novel features of communities not simply accessible through other means.

(b) *MOTU inference using different cutoff scores*

We performed 300 independent clusterings of the 295 tardigrade nSSU sequences. One hundred independent, random-addition order replicates were produced for taxon definition cutoffs of 2, 3 and 4 bp. For the 2 bp cutoff, the number of MOTU inferred ranged from 143 to 157, with a mode of 151 and a mean of 149.96 ± 2.61. The majority of the variability in MOTU number inferred was due to alternate groupings of a few clouds of sequences (not shown). The use of larger cutoff values also resulted in MOTU sets with wide ranges ($\sim$10% of the total number inferred) (figure 5). Thus increasing the fuzziness of the MOTU discriminant does not result in a simple collapse of the clouds of sequences into single taxa. We have also observed this pattern in other meiofaunal datasets (Blaxter *et al.* 2003) (Floyd, Blaxter *et al.* unpublished).

The variability of attribution observed between independent clusterings is not a unique feature of MOTU_define.pl: the same issue must arise in all other methods, as the data we are using to infer taxa is essentially quantal.

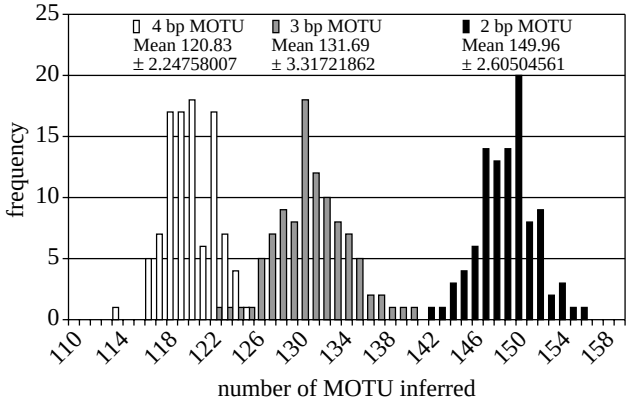

Figure 5. Variability in the number of MOTU defined by replicate analyses. The histogram shows the frequency distribution of total numbers of clusters inferred from 295 tardigrade nSSU sequences from the Glen Ettrick study site using MOTU_define.pl at three different cut off values: 2 bp (black), 3 bp (hatched) and 4 bp (open). The mean and standard deviation of each set of analyses is given.

6. DISCUSSION: TURNING SEQUENCES INTO MOTU

The MOTU-define.pl program is but one, obvious way of inferring MOTU. Other methods could also be applied. One common choice is to use a BLAST algorithm (usually BLASTn) (Altschul *et al.* 1997) to identify the best match in a reference database, and to assign the identifier of the best match to the barcoded specimen if the similarity is judged to be good enough. This method has many pitfalls, not least its reliance on a well-populated (and correctly named) database of barcodes. In meiofaunal surveys such as presented here, the lack of close relatives in the database can make this approach less-than-rewarding. More importantly, the BLAST algorithm (Altschul *et al.* 1990; Altschul *et al.* 1997) was not designed for barcode identity assignment, and simply taking the top-scoring match, with some predefined quality score cutoff, may miss issues of, for example, equal top scoring matches. A variation on the BLAST approach would be to extract the best matches (for example, all matches with a score within a small percentage of the best match), perform a complete alignment with the barcode query sequence, and then subject this alignment to model-driven phylogenetic analysis to ask if the barcoded specimen is a credible member of a monophyletic clade with any of the references.

Because much DNA barcode sequence is derived from single sequencing reads on only one strand of the DNA, the quality of the sequences may not be as good as those in the databases. The sequencing chromatogram can be analysed to yield a quality score for each base (Ewing & Green 1998; Ewing *et al.* 1998), and these could be incorporated into a BLAST-and-align method for MOTU definition that down-weights any differences associated with low quality scores and pays more attention to high-quality scores. A variation on this method might also include partitioning the aligned

sequences *a priori* into more- and less-informative sites. Thus, in a protein-coding gene such as *cox1*, one might give first and second base changes more weight than those observed in fourfold degenerate sites. In a RNA gene such as nSSU, one could differentially weight residues by their involvement in secondary structure, and their observed conservation in large aligned datasets.

As barcoding is applied somewhere on the span between population genetics and taxon phylogenetics, the use of network-based algorithms may also assist. Templeton network analysis is much used in population studies to examine patterns of haplotype distribution and relatedness (Clement *et al*. 2000). For DNA barcode data, such network analysis, with different cutoffs for the breaking of ties between subnetworks, can assist in understanding the patterns of diversity in the sequences and thus the likely status of the MOTU defined. In genomics, definition of protein families has been achieved using multiple cluster linkage methods, where complex networks of similarity between sequences can be examined at different levels of granularity to identify coherent clusters (Enright *et al*. 2002). A similar approach applied to DNA barcode data might be doubly informative of not only final MOTU but also the interrelationships of MOTU clouds.

Ultimately, we might want to use rigorous phylogenetic methods to affirm the monophyly of our newly defined MOTU, and to place them in the context of named sequence diversity. However, we must be aware of the issues of partial sorting of haplotypes between lineages as they diverge. Wide-ranging studies on several taxa have clearly shown that while rapidly evolving sequences are very well suited to generation and testing of taxon hypotheses at local scales, they are often very much unsuited to deeper phylogenetic analysis. Processes such as base substitution bias and variable site saturation can rapidly obscure real phylogenetic signals and generate spurious trends in data. The barcode data will be rather unsuitable for reconstructing the deeper branches of the tree of life, including perhaps all those below the generic level (Vogler *et al*. 2005). Simply using trees to infer taxa from barcode data can be positively misleading: we should rather define the taxa and then examine their relationships through rigorous phylogenetics.

Taxa defined by MOTU methods can be used for standard taxonomic and ecological surveys. By comparing the barcode sequence with a database of sequences from specimens identified to Linnaean taxa before sequencing, the anonymous survey specimens can be placed within the known taxonomic framework, and the organismal biology of the organisms from which they derived inferred (Floyd *et al*. 2002; Blaxter & Floyd 2003; Blaxter 2004). By this method we can move from anonymous sequence to ecosystem biology.

This work was carried out as part of ongoing investigations into meiofaunal diversity in our laboratory, and was funded by the UK Natural Environment Research Council and the Linnaean Society of London. J.M. and T.C. carried out the meiofaunal surveys as part of their major undergraduate projects.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Evol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389.)

Amaral Zettler, L. A., Gomez, F., Zettler, E., Keenan, B. G., Amils, R. & Sogin, M. L. 2002 Microbiology: eukaryotic diversity in Spain's River of Fire. *Nature* **417**, 137. (doi:10.1038/417137a.)

Berry, W. 1970. *Farming: a handbook*. San Diego, CA: Harcourt Brace & Company.

Blaxter, M. 2003 Molecular systematics: counting angels with DNA. *Nature* **421**, 122–124. (doi:10.1038/421122a.)

Blaxter, M. L. 2004 The promise of a molecular taxonomy. *Phil. Trans. R. Soc. B* **359**, 669–679. (doi:10.1098/rstb.2003.1447.)

Blaxter, M. L. & Floyd, R. 2003 Molecular taxonomics for biodiversity surveys: already a reality. *Trends Ecol. Evol.* **18**, 268–269. (doi:10.1016/S0169-5347(03)00102-2.)

Blaxter, M., Elsworth, B. & Daub, J. 2003 DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. *Biol. Lett.* **271**, S189–S192. (doi:10.1098/rsbl.2003.0130.)

Blaxter, M., Floyd, R., Dorris, M., Eyualem, A. & De Ley, P. 2004 Utilising the new nematode phylogeny for studies of parasitism and diversity. In *Nematology monographs and perspectives* (ed. R. Cook & D. J. Hunt), pp. 615–632. Leiden: E. J. Brill.

Blaxter, M. L. *et al*. 1998 A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75. (doi:10.1038/32160.)

Clement, M., Posada, D. & Crandall, K. A. 2000 TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1659. (doi:10.1046/j.1365-294x.2000.01020.x.)

De Ley, P. & Bert, W. 2001 Video capture and editing as a tool for storage, distribution and illustration of morphological characters of nematodes. *J. Nematol.* **34**, 296–302.

De Ley, P. *et al*. 2005 An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Phil. Trans. R. Soc. B* **360**, 1945–1958. (doi:10.1098/rstb.2005.1726.)

Diez, B., Pedros-Alio, C. & Massana, R. 2001 Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941. (doi:10.1128/AEM.67.7.2932-2941.2001.)

Enright, A. J., Van Dongen, S. & Ouzounis, C. A. 2002 An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584. (doi:10.1093/nar/30.7.1575.)

Ewing, B. & Green, P. 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.

Eyualem, A. & Blaxter, M. 2003 Comparison of biological, molecular and morphological methods of species identification in a set of cultured *Panagrolaimus* isolates. *J. Nematol.* **35**, 119–128.

Finlay, B. J. 2002 Global dispersal of free-living microbial eukaryote species. *Science* **296**, 1061–1063. (doi:10.1126/science.1070710.)

Floyd, R., Eyualem, A., Papert, A. & Blaxter, M. 2002 Molecular barcodes for soil nematode identification.

Mol. Ecol. **11**, 839–850. (doi:10.1046/j.1365-294X.2002.01485.x.)

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. 1990 Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63. (doi:10.1038/345060a0.)

Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218.)

Kinchin, I. M. 1994. *The biology of tardigrades*. London: Portland Press.

Lambshead, P. J. D. 1993 Recent developments in marine benthic biodiversity research. *Oceanis* **19**, 5–24.

Lambshead, P. J. D. & Boucher, G. 2003 Marine nematode deep-sea biodiversity—hyperdiverse or hype? *J. Biogeo.* **30**, 475–485.

Lawton, J. H. *et al.* 1998 Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forests. *Nature* **391**, 72–75. (doi:10.1038/34166.)

Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D. 2001 Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607.

Maucci, W. (ed.) 1986 *Fauna d'Italia Tardigrada*. Bologna: Edizioni Calderini Bologna.

May, R. M. 1988 How many species are there on earth? *Science* **241**, 1441–1449.

Moreira, D. & Lopez-Garcia, P. 2002 The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* **10**, 31–38. (doi:10.1016/S0966-842X(01)02257-0.)

Parkinson, J., Guiliano, D. & Blaxter, M. 2002 Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics* **3**, 31. (doi:10.1186/1471-2105-3-31.)

Platt, H. M. 1994 Foreword. In *The phylogenetic systematics of free-living nematodes* (ed. S. Lorenzen). London: The Ray Society.

Swofford, D. 1999. *PAUP** 4.b10. Sunderland, MA, USA: Sinauer Associates.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74. (doi:10.1016/S0169-5347(02)00041-1.)

Vogler, A., Cardoso, A. & Barraclough, T. 2005 Exploring rate variation among and within sites in a densely sampled tree: species level phylogenetics of north american tiger beetles (genus *Cicindela*). *Syst. Biol.* **54**, 4–20.